Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 13:29:08
REFERENCE NO: 243

Author Names & Affiliations

- John Williams - University of Wisconsin-Madison
- Simon Goring - University of Wisconsin-Madison
- Julien Emile-Geay - University of Southern California
- Douglas Fils - Consortium for Ocean Leadership
- Eric Grimm - University of Minnesota
- Kerstin Lehnert - Columbia University
- Nick McKay - Northern Arizona University
- Amy Myrbo - University of Minnesota
- Anders Noren - University of Minnesota
- Lisa Park Boush - University of Connecticut
- Shanan Peters - University of Wisconsin-Madison
- Brad Singer - University of Wisconsin-Madison
- Mark Uhen - George Mason University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Earth System Science; Paleogeoscience: geochronology, geoinformatics, paleoclimatology, paleoecology, paleobiology, sedimentology, stratigraphy

Title of Submission

Cyberinfrastructure in the Paleosciences: Mobilizing Long-Tail Data, Building Distributed Community Infrastructure, Empowering Individual Geoscientists

Abstract (maximum ~200 words).

In an era of global change, we use paleogeoscientific data to study how the Earth-Life system responds to and recovers from large perturbations to the global carbon cycle, biodiversity, climates, cryosphere, and hydrosphere. The grand informatics challenge is to organize and mobilize billions of observations distributed across space, time, disciplines, and institutions, so that we can bring all relevant data to bear on any time, place, or process. The emerging cyberinfrastructure model consists of a distributed, federated network of resources, with community curated data repositories (CCDRs), physical sample repositories, individual geoscientists, the scientific literature, and networking/coordination efforts. In our field, the most productive scientific return from NSF cyberinfrastructure investments will come from distributed, meso-scale investments: 1) Long-term investments in the human capital necessary to develop and sustain community-curated data resources (CCDRs), 2) Data mobilization campaigns targeted to high-priority research questions, 3) Scientific workforce training at all career stages, 4) Reduced data friction via integrated data handling systems from field collection to measurement,

paper publication, and data publication, 5) Automated data-mining systems for extracting information from unstructured sources, 6) A National Center for Paleodata Synthesis to accelerate and coordinate the above global-scale science and informatic activities.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

1.1 Paleogeosciences: Major Scientific Questions and Research Challenges
The grand challenge in the paleogeosciences is to enable a fully resolved understanding of the past dynamics of the Earth-Life System and its interacting subsystems, across the entire history of Earth, at temporal scales of 109 to 101 years, by organizing and mobilizing the many millions of individual geoscientific observations that make up the long tail of paleogeoscience data (Transitions Report, 2012, Earth Cube Paleogeosciences Domain Workshop 2012, NRC 2013, 2011a,b). The paleogeosciences branch of Earth System Science encompasses paleoclimatology, paleobiology, paleoecology, geochronology, sedimentary geology, geochemistry, glaciology, and other disciplines. In our era of global change, with projected rates of change and states of the climate system with no analog in recorded human history, the paleogeosciences are vital to studying how the Earth-Life system responds to and recovers from large perturbations to the global carbon cycle, global biodiversity, regional and global climates, cryosphere, and hydrosphere.
Four overarching scientific challenges in Earth System Science were identified in the National Research Council's Transitions Report (2012):
*What is the full range of potential climate system states and rates of transitions experienced on Earth?
*What are the thresholds, feedbacks, and tipping points in the climate system, and how do they vary among different climate states?
*What are the ranges and rates of ecosystem response, modes of vulnerability, and resilience to change in different Earth system states?
*How have climate, the oceans, the Earth's sedimentary crust, carbon sinks and soils, and life itself evolved together, and what does this tell us about the future trajectory of the integrated Earth-Life system?

See also National Research Council reports (2013, 2011a,b) and EarthCube Domain Working Group Reports (Noren et al. 2013, Aufdenkampe et al. 2013, Chan and Budd 2013, and Singer et al. 2013, all ref: http://bit.ly/2nUOUQc).
We can answer these questions through the study of Earth's history and its rich record of past abrupt change, evolutionary innovations, and complex dynamics driven by interactions among multiple components of the earth system, across multiple temporal and spatial scales. Earth's history provides multiple model systems for 21st-century changes (Williams et al. 2013). Areas of active research include:

*The effect of early life on atmospheric evolution and global geochemical cycles (e.g., Peters et al. 2017).
*The five major mass extinctions of Earth's biosphere and understanding the processes that govern rates of speciation and evolutionary innovation.
*Disruptions to the Earth's global carbon cycle, (e.g., the rapid release of organic carbon into the atmospheric-ocean system during the Paleocene-Eocene Thermal Maximum, and ensuing effects on species extinction and evolution).
*The glacial-interglacial cycles of the Quaternary, paced by variations in the Earth's orbit and amplified by feedbacks among ice sheet dynamics, ocean circulation and chemistry, and climate.
*The persistence of biodiversity during past glacial periods and the community turnover and species range shifts during past glacial-interglacial cycles
*Reconstructing global and regional temperature trends over the current interglacial and last millennium, while disentangling the effects of external forcings, internal feedbacks, unforced variations, and the growing anthropogenic footprint.

1.2 Paleogeoscientific Data: Key Features and Challenges
Here we summarize key features of paleogeoscientific data, practice, and practitioners. These characteristics have been the starting point for current cyberinfrastructure-building efforts (Sect.2.1) and inform our recommendations for the next generation of cyberinfrastructure advances (Sect.2.2-3.5).
1. Paleogeoscientific observations are long-tail data collected by scientists from many disciplines and institutions, with many data types and forms of measurement. Individual records are temporally long but spatially point-level data, collected at one or more outcrops, drill sites, or other discrete sites. Hence, site-level paleogeoscientific data must be assembled into global-scale data networks in order to understand the Earth System, its external forcings, and internal feedbacks (e.g. PAGES 2k, 2013). Assembling such data is labor-intensive. Few widely accepted data standards and identifiers exist (McKay & Emile-Geay, 2016), although several are emerging through EarthCube-supported Research Coordination Networks (Cyber4Paleo: https://www.earthcube.org/group/c4p) and Integrative Activities (ePPANDA, Earth-Life Consortium, Open Core Data).

2. Paleogeoscientific data share common underlying structure. Despite the above heterogeneity, paleogeoscientific data share several underlying common features: They typically involve a measurement of a proxy in various geological archives, often structured by depth, from which we must estimate time. This structural homogeneity facilitates the development of common data models in the paleogeosciences.
3. Paleogeoscientific data has a long shelf life. Paleogeoscientific data derive primarily from physical samples of geological materials collected in the field and the laboratory measurements of these samples. As new techniques are developed, we often seek to reanalyze previously collected samples, cf. the recent wave of ancient DNA analyses from museum fossils. We must curate physical samples and maintain an unbroken chain of provenance from sample to all data generated from the sample (Sect.2.3).
4. Time is an unknown variable that must be estimated in the paleogeosciences (Singer et al. 2013). We must infer age through discrete age estimates (called age controls) and age models that provide age estimates between dated samples. Age models must be regularly updated as more precise and accurate dates become available and as more sophisticated age-depth software modeling approaches are developed. Published geochronological frameworks become obsolete with every new date and refinement to dating methods, decay constants, and other parameters. Data repositories exist for some geochronological data (GeoChron/IEDA: http://www.geochron.org), but they are not systematically linked to one another or to other affiliated databases.
5. Dark Data. Data are often not fully published. For example, papers presenting microfossil data often show only summary diagrams for selected taxa and may fail to include supplementary data. Published metadata are incomplete, e.g. geochronological labs usually do not publish all instrumental parameter settings. Some disciplines have adopted minimal metadata standards and established a common data repository; others have not. A great deal of data is still digitally “dark”, even if publications themselves are available electronically. Data mobilization efforts are essential (Sect.3.3).
6. Paleodata are increasingly assimilated with Earth System Models. Our field uses Earth system models to simulate the processes governing the past and present evolution of the Earth-Life system. These same models are also the basis for climate scenarios over the coming decades, and paleodata offer an important constraint on modeled estimates (e.g. sensitivity of global temperatures to atmospheric CO2, Hargreaves et al. 2012). Increasingly, data assimilation methods are being employed to make joint inferences from paleodata and Earth system models (Crucifix, 2012). For example, atmospheric general circulation models now include stable isotopic tracers (e.g. d18O), enabling direct assimilation of earth system models with paleodata. Data assimilation creates new needs for well-structured datasets with rigorous estimates of temporal and proxy uncertainty and for high-capacity computing.
7. Paleogeoscientific Expertise is Widely Distributed, with individual paleogeoscientists specializing in particular proxy types, archives, time periods, regions, and questions. Dispersion of expertise places a premium on developing decentralized, but interlinked governance and data management systems for our data (Sect.2.1, 3.1-3.2)
8. Uneven Workforce Training and Interest in Informatics. The paleogeosciences emphasize high-quality field and laboratory measurements. Informatics has not traditionally been part of the core geoscientific curriculum, except for courses in statistics and calculus. Most geoscientists have not sought to keep pace with recent rapid advances in informatics. Disciplinary and cultural norms vary with respect to data sharing. Training programs at all levels are needed (Sect.3.4).

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

2.1 Cyberinfrastructure in the Paleogeosciences: Trends and Needs
In the paleogeosciences, the emerging cyberinfrastructure model is a distributed, federated network of resources and services, each curating and advancing a particular kind of knowledge. This system is consistent with the distributed nature of geoscientific expertise: geochronological data should be curated by geochronologists, species names by paleontologists, etc. Within this structure, focal nodes serve as disciplinary loci for data stewardship and mobilization, and as resources for sharing best practices and standards across subdisciplines. This structure has emerged organically, with key support coming from NSF Geoinformatics and EarthCube, and at least five major interacting components:
Community Curated Data Repositories (CCDRs). CCDRs have repeatedly emerged as geoscientists unite to gather and share data in response to common, broad-scale research questions (Section 1A). CCDRs usually begin as individual or small-team efforts, then mature into community resources with established data models, standards, and governance systems. Because describing the history of Earth’s climates and biodiversity requires coordinated effort, CCDRs prevail in paleoclimatology, paleobiology, and paleoecology (e.g., Neotoma Paleoecology DB, Paleobiology Database, LinkedEarth, SedDB/EarthChem, NOAA Paleoclimatology, MorphoBank, VertNet, MorphoBank). New CCDRs emerge as proxies mature, e.g. the recent call for IsoBank (Pauli et al. 2017). Key challenges include reducing data friction by

better integration (Sect.2.3), encouraging community input, and, most of all, sustainability (Sect.3.1).

Museums and Sample Repositories curate physical specimens (rock samples, drill cores, fossils, biological materials) and their digital representations. Examples include LacCore/CSDCO, IODP, and marine core repositories at Columbia University, University of Rhode Island, and Oregon State. Recent initiatives have focused on digitizing collections (iDigBio, iDigPaleo), developing as persistent and unique sample identifiers (IGSNs), and establishing provenance systems for linking samples to measurements (Open Core Data).

Integration and Networking Activities. The EarthCube initiative has provided an essential push and resources to build a networked federation of interconnected CCDRs and sample repositories. Current efforts include Open Core Data (linked data standards for continental and ocean marine drilling data), the Earth Life Consortium (an umbrella organization for Neotoma, PBDB, and other paleobiological CCDRs, and linking to modern biodiversity databases), ePANDDA (linking the specimen holdings in Paleobiology Database with the museum digitization efforts by iDigPaleo), LinkedEarth (linked data standard for paleoclimatic data, McKay et al. 2016), and Flyover Country (a popular app for viewing geological data during air or ground travel, now used by ~600 people each day for informal learning and a recent “Vizzie” award).

Individual geoscientists. Most scientific data curation is still done by individual geoscientists in their research labs, with data stored on desktop computers and local servers. Most geoscientists store and record data using flat file formats (TXT, CSV, XLS) or workflow software associated with their instrumentation systems. Many subdisciplines have no established metadata standards, data reduction standards, or community data repositories. Huge effort is spent converting small amounts of data from one format to another (by some estimates, up to 80% of total project effort).

Scientific Literature and Unstructured Data. A vast amount of paleogeoscientific data is available only through the published literature. These data are highly unstructured and not readily amenable to broad-scale synthesis. Arguably, over 100 years of geological research has mainly succeeded in transferring information from one vast and dimly accessible archive (the geologic record) to another (the published literature — better, but far from ideal). We need better systems for mining this resource.

2.2 Paleogeoscience Cyberinfrastructure: Priorities for the Next Decade

Based on the above, we argue that for our community, the most productive scientific return on NSF cyberinfrastructure investments over the next decade will come from distributed, meso-scale investments, with the following six priority areas:

1. Reduce data friction by developing scientific workflows, structured vocabularies, semantic frameworks, and data-tagging systems to pass data and metadata seamlessly within and among community resources. (Sect.2.3)
2. Develop automated data-mining systems for extracting information from unstructured data in the scientific literature (Sect.2.4).
3. Support the long-term sustainability of existing community cyberinfrastructure resources (Sect.3.1) and the grassroots development of community informatics resources for sub-disciplines that lack data sharing systems. (Sect.3.2).
4. Launch funded data mobilization campaigns to unlock existing data relevant to high-priority scientific research questions (Sect.3.3).
5. Develop and train a distributed scientific workforce, for both early career scientists and current practitioners (Sect.3.4).
6. Establish a National Center for Paleodata Synthesis to coordinate activities among individual geoscientists and the federation of CCDRs and sample repositories, promote community best-practices and data standards, and develop education and scientific workforce training initiatives (Sect 3.5).

2.3 Reduce Data Friction

We need to build the systems and standards necessary to pass data within primary data-generation scientific workflows (from field collection to laboratory measurement to publication and archival) and among the emerging federation of data repositories that facilitate downstream data integration and synthesis. Researchers must be able to access data from any point in the stream of data generation, and see its provenance, the implications or effects of models on their data, and its subsequent interpretations. Repositories must be interconnected, and scientists must have the ability to iteratively re-evaluate and annotate data. Clear provenance is essential to link records across resources.

Web architectural approaches are federated, scalable and tolerant; all are desirable properties for a distributed paleodata network. Common permanent and persistent identifiers such as DOIs, IGSNs and ORCIDs, combined with linked open data and semantic frameworks, are needed to enable the passing of data among community supported resources. The development and inclusion of well-documented data vocabularies would improve the ability to move across data repositories and disciplinary divides, clearly identify the assumptions surrounding data objects, and simplify translational activities, supporting the development of the semantic web. In particular, the development of an ontology of geologic time (e.g., OWL Time) should be a priority. These can then be extended into new and underserved data communities, accelerating the mobilization of large volumes of long tail data.

Several national and international efforts in this area are already underway, e.g. RDA, ESIP, EarthCube, and Mozilla. We need mechanisms for better sustained engagements between these efforts and paleogeoscientists (Sect.3.5).

2.4 Machine Reading Systems to Facilitate Creation of Structured Knowledge Bases From Unstructured Data

The ability to algorithmically and repeatedly interrogate the scientific literature, en masse, for the purpose of locating and extracting the data needed to address broad-scale research questions would revolutionize efforts towards building a fully realized, data-constrained model of the evolving Earth-Life system. Current efforts to aggregate, organize, and synthesize paleogeoscientific data rely heavily on manual literature-based data compilation, which is labor-intensive, costly, and rate-limiting.

Machine reading and learning systems are rapidly advancing and hold promise as scientific research tools for literature-based data compilation efforts (Ré et al. 2014; Mallory et al. 2015; De Sa et al. 2016a, b; Peters et al. 2014, 2017). EarthCube's GeoDeepDive is a major step in this direction, with a digital library of over 3 million documents from multiple partners and >1 million CPU hours invested in parsing and annotating these documents. The next step is to grow and build all-inclusive digital libraries of published scientific documents and the associated high-capacity computing infrastructure that enables scientists to search the literature and dynamically create structured research syntheses.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

As practicing geoscientists who work at the science-informatics interface, it is our firm belief that the critical barrier to big-data science in the paleogeosciences is not technology, it is people. Sustained investment in human capital is essential, and has several dimensions.

3.1 Sustainability

Sustained Support for Community Cyberinfrastructure. NSF has an excellent system for seeding and building cyberinfrastructure (funded workshops to build initial connections, RCNs to develop networks, 3-year grants to develop resources), but NSF lacks a good system for long-term sustenance of paleodata cyberinfrastructure. Many of the cyberinfrastructure resources (PBDB, Neotoma, LacCore, and precursors) mentioned here have existed for decades and have achieved sustainability by being closely linked to research priorities. But their existence is always precarious and depends on funding in 3-year grant cycles, which is a major barrier to bringing in new data and colleagues. When we ask colleagues to participate in these initiatives, by contributing data or lending expertise, a question that always comes up is sustainability. Meaning, 'as a hard-working geoscientist, with many demands on my time, should I invest any effort in assisting with a resource that may not exist in three years'?

The lack of long-term funding is incongruous with CCDRs' increasing role as data repositories for NSF-sponsored research, thereby fulfilling Federal mandates for public data sharing. Cyberinfrastructure is infrastructure, and should be supported with long-term (5-10yr) investments, contingent upon satisfactory support of community research priorities.

Sustainability: Human Capital. Much infrastructure focuses on 'hard-capital' resources: ships, planes, satellites, core repositories, etc. In cyberinfrastructure, human capital is paramount, and comprises (1) trained domain experts for data acquisition and quality control, and (2) IT experts who design and maintain databases and the software interfaces for data entry and retrieval. Through hard experience, we have learned that exceedingly few experts have the necessary joint training in the geosciences and data sciences. Our data are complex, with substantial embedded knowledge. Our successes and advances have depended on a few individuals who have, through on-the-job work, acquired the necessary crossover training.

In drafting this letter, one contributor [Emile-Geay] wrote: "Right now, I have an outstanding postdoc who is doing more useful work for our project than 5 technicians… She'd love to continue working for the LinkedEarth project, but the lack of sustained funding means that she will probably have to move on." Another [Peters] noted: "My guy is brilliant for PBDB/Macrostrat/GeoDeepDive. But he didn't 'come that way.' He has a BA in Political Science. He… acquired experience of high value on this job with my group.“

The technology associated with much paleogeoscientific cyberinfrastructure is low-cost and resilient to funding interruptions – servers, APIs, etc. It is the loss of key individuals and their embedded knowledge that cripples cyberinfrastructure initiatives. We are continually at risk of losing talented data scientists because of funding lapses or low pay relative to skill sets. Academic salaries are much lower than industry salaries for many technical staff. Without adequate reward of these mission-critical crossover data scientists and geoscientists, with deep disciplinary knowledge and key cyberinfrastructure skills, we risk losing irreplaceable expertise and the sustainability of current cyberinfrastructure efforts.

3.2 Bottom-Up Development of Community Informatics Resources:

Many disciplines do not yet have agreed-upon CCDRs or minimal metadata standards. These communities should be encouraged to self-organize through the established NSF mechanisms of workshop grants, RCNs, and seed grants. The new EarthRates RCN is a good step in this direction. Emerging initiatives should be partnered with established initiatives, to minimize reinventions of wheels and to encourage adoption of common best practices. Clear documentation and support for domain scientists must be generated to bridge disciplinary gaps, particularly within fields with less of a background in informatics. Clear assessment guidelines for new developments must be supported, so that new infrastructure or data models can be evaluated, and best practices can be embedded at the earliest development stages.

3.3 Data mobilization and improvement campaigns.

Data mobilization campaigns are vital, given the prevalence of dark and incompletely published data (Sect 1.2). We need to provide resources to people to upload their data from in-house computers to community databases.

In the paleogeosciences, the same pattern repeats over and over: A global data synthesis is launched to target a particular time interval or broad-scale scientific qeustion, often with workshop support from NSF or PAGES (e.g. Climates of the last two millennia; Pliocene data-model syntheses). Scientists contribute their individual data files. Conveners and contributors quickly discover massive heterogeneity in the individual spreadsheet datafiles. The project stalls, scales back ambitions, or takes years to complete. After publication, results are often not readily re-usable because few resources were invested in proper data publication.

We need a new model in which research synthesis projects are explicitly combined with data mobilization campaigns. Research teams should apply for data mobilization funds, in which they identify a critical scientific problem that would benefit from mobilization and synthesis of existing data. Funding should include workshop support and support for postdocs, grad students, or technicians to receive data from participants and upload to community databases conforming to recognized standards. These funds could be awarded and workshops run through NSF or a designated synthesis center (Sect 3.5). Data mobilization efforts should prioritize the foundational ‘raw’ data (e.g. radiocarbon dates, geochemical measurements, fossil occurrences) and secondarily the ‘derived’ data (e.g. age models, temperature reconstructions, etc.) that source from the raw data. The paleogeosciences could adopt the EarthScope terminology of Level 0 (raw, unprocessed), Level 1 (quality-controlled data), Level 2 (low-level derived products), Level 3 (mid-level integrated products), Level 4 (high-level integrated products) (http://www.usarray.org/files/docs/pubs/ES_Data_Portal.pdf) and prioritize data mobilization from the bottom up.

3.4 Scientific Workforce Development

We need targeted initiatives to better train our scientific workforce in best practices in data handling and synthesis. This includes community development platforms such as GitHub, scientific workflow systems, and efforts towards transparent, reproducible science. Training is needed at all levels, including undergraduate, graduate, and refresher training for early-career and mid-career scientists. Delivery options include IGERT-style graduate training programs, YouTube, and summer workshops, e.g. software carpentry (https://software-carpentry.org/) and data hackathons (http://cyber4paleo.github.io). We need to rebuild undergraduate and graduate courses in the geosciences to emphasize data science, with more emphasis on scientific programming and coding practices, hierarchical Bayesian statistics, and geovisualization. Our ultimate goal should be to build the next generation of geoscientists who transform science through their work at the science/informatics interface.

3.5 National Centers for Paleodata and Synthesis (NCPDS)

A key idea in all of the above is distributed. It would be a grave mistake to try to create a single highly centralized data center in the paleogeosciences. Given the heterogeneity of our data and dispersal of communities, we envision a federated ecosystem of resources, each serving their respective community with sustained support and integrated management. Management should be through a coordinating office that facilitates standards adoption, provenancing, and other tools for data sharing across CCDRs, promotes community best practices, and leads scientific workforce training initiatives. This Center would help facilitate connections to other existing organizations such as ESIP, Mozilla Science, ICSU, and coordinate activities with EarthCube and NSF directorates. Possible models include the coordinating office for the Long-Term Ecological Research (LTER) Network, the NSF Centers for Synthesis (NCEAS, NESCENT, SESYNC), and the USGS Powell Center (https://powellcenter.usgs.gov/).

Consent Statement